Lec20 MATH 343

We usually don't know when b_j is a causal estimate or not.

However, at the very least, it is correlational.

What is the interpretation of a correlational estimate?

Warning: it's weak! Here it is for OLS, x_j = blood sugar (mg/dL), y = blood triglycerides (mmol/L) for $b_j = 1.17$, $s_{b_j} = 0.25$

"When comparing two observations (A) and (B) which are sampled in the same fashion as the observations in $\mathbb{D}$ were sampled when (A) has blood sugar (name of metric x_j) 1 mg/dL (unit) larger than (B) but share the same measurement values otherwise, then (A) is predicted to have an estimated blood triglycerides level (name of response y) 1.17 ± 0.25 (value of $b_j \pm s_{b_j}$) mmol/L (unit of y) higher (if $b_j > 0 \Rightarrow$ higher o/w lower) than (B)'s blood triglycerides level (name of response y) assuming the blood triglycerides level is linear in the p covariates. (the response function assumption)

do logistic regression, weibull survival, cox ph.

For prediction only $\hat{y} = g(\vec{x}_*)$, we don't care at all about causation! We just want the features to have non-spurious correlations with y! This is why this topic was excised from 342. Because it's irrelevant!!

Why is the causal interpretation more powerful than the correlational interpretation? Causal relationship means if you do the manipulation, you get the results. You don't get this with a correlational relationship!! Agenda for prediction, is doesn't matter!

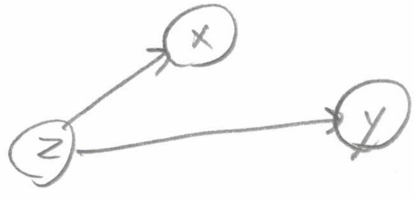
Prediction answers question: what will happen? "You will make more money if x increases naturally"

Causality answers question: what will make it happen?

"You will make more money if you increase x."

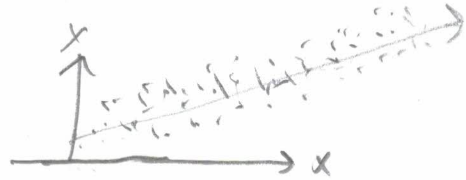
Causality is big business because everyone wants to know what they can do to make something happen.

When/why exactly do correlations fail to be causal. Take [B]



x = # of umbrellas sold 6-9AM in NYC
y = # of car accidents 9AM-9PM in NYC
z = # inches rain forecast in NYC at 6AM

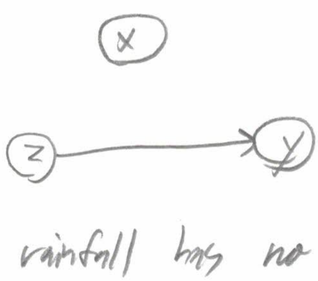
Obviously Corr[x, y] ≠ 0 and positive



b_1 would be positive and significant

You cannot use this scatterplot as a means to assess causality, you must have other information! You must have the DAG and it must be true!

Let's say the govt goes in and manipulates umbrella sales. They say, you cannot sell to NYC residents, you only sell to me so I can ship them to Paraguay. Then they proceed to give them varying volume orders every day. By manipulating x, it is now completely disconnected from arrows coming in:



Thus rainfall has no effect on x, hence there is no information about y in x.

$b_1 \approx 0$ and insignificant

This constitutes proof x is <u>not</u> causally related to y!
We will return to manipulation later!

So far, we've only been examining simple regressions.

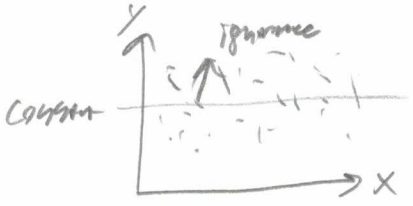
What happens if we run the regression and fit $\hat{y} = b_0 + b_1 x + b_2 z$?

What is interpretation of b_1? "...but slave the same measurement values otherwise...". This means consider two observations A, B

$x_A = x_0$, $x_B = x_0 + 1$ where $z_A = z_0$ and $z_B = z_0$. What is the difference in the expected response?

Linear regression is a way of looking at a conditional universe!

Consider $z_A = 17$ and $z_B = 17$. The world of all observations $\langle x, 17, y \rangle$ looks like



$b_1 \approx 0$ and insignificant! Why? We're conditioned on rainfall. So $y = k_{y,z}(17) + \text{ignorance}$
$= \text{constant} + \text{ignorance}$

This is what b_1 in the regression $\hat{y} = b_0 + b_1 x + b_2 z$ is estimating, β_1

It's completely different from β_1 estimated in $\hat{y} = b_0 + b_1 x$. DEMO

It's a change in perspective, not universe itself. Manipulation: change in universe itself.

This seems unfair! If we add a feature to a regression does estimate change? Go from positive and significant to zero and insignificant?

Yes! They can do anything. It's all based on the underlying structure given by the DAG.

Paradoxes!

You may be thinking, this umbrella-car-accidents with lurking variable main is obvious. I'll never fall into the trap the correlation $\Rightarrow$ causation. Not so fast. Consider

x_1 = exercise amount and y = cholesterol level. The data shows:

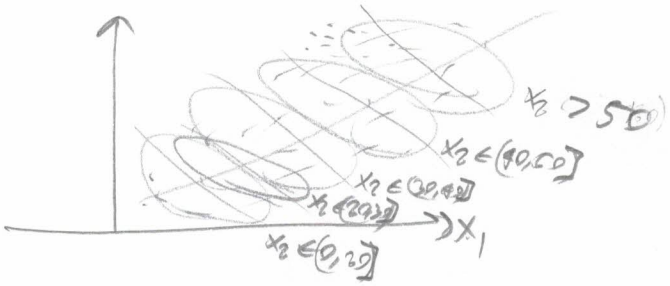
$\hat{y} = b_0 + b_1 x_1 \Rightarrow$

$b_1 > 0$ and significant

the data is real!

You're taught to say more exercise causes higher cholesterol.

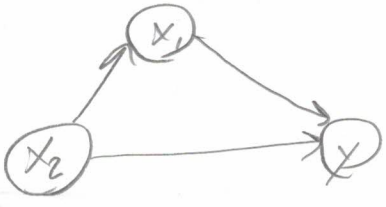
How did this happen? There's another variable, x_2 = age.



This is called "Simpson's Paradox" (1951)

If you condition on x_2 i.e. real $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$, then you reveal the real causal effect. DAG? Nevermind...

D



$\exists$ a formal

$\Rightarrow$ The effect of x_1 on y is causal if x_1 drives causation and if there is additionally a common cause, it's controlled for. (You need Multivariate Regression in the latter case.